DOCUMENT RESUME

ED 152 850 TM 007 068

AUTHOR Linehan, Marsha M.
TITLE Content Validity in Behavioral Assessment.
PUB DATE [Sep 76]
NOTE 13p.; Paper presented at the Annual Meeting of the American Psychological Association (84th, Washington, D.C., September 3-7, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Behavior Patterns; Content Analysis; *Criterion Referenced Tests; Informal Assessment; *Item Sampling; Measurement Techniques; *Observation; Test Construction; Test Reliability; *Test Validity
IDENTIFIERS *Behavioral Assessment; *Content Validity; Idiographic Approach; Nomothetic Approach

ABSTRACT

Both criterion-referenced testing and behavioral assessment share the basic assumption that test behavior is a sample rather than a sign. In addition, both types of assessment focus on response capabilities and performance in specified content domains. Although content validity has been traditionally recognized as essential to criterion-referenced testing, it has not been as generally recognized as relevant to behavioral assessment. The relationship between the two types of assessment clarifies the importance of content validity for behavioral assessment. Questions of content validity can be usefully conceptualized from the point of view of generalizability theory. The relevant question then becomes: Are the test items a representative sample of the universe to which results are going to be generalized? Finally, content validity is as important in assessment procedures developed on an idiographic level as it is in those developed nomothetically. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED152850

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Marsha Linehan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Content Validity in Behavioral Assessment

Marsha M. Linehan
The Catholic University of America

Paper presented at the
American Psycological
Association Meeting
Washington, D.C.,
1976

TM007 068

CONTENT VALIDITY IN BEHAVIORAL ASSESSMENT

Historically, behavioral assessors have criticized traditional assessment methodologies on the grounds of limited validity and utility (Mischel, 1968). A frequent response to that critique has been a call for behaviorists to develop assessment methods to replace the traditional ones. At least partially in response to this demand, there has been an explosion in the last several years of published behavioral assessment methodologies, tests formulated within a behavioral framework, and proposed applications of behavioral testing procedures. There has not, however, until recently been a corresponding interest in examining the relevant validities of behavioral assessment methods. In part, this has been because of the notion that if the criterion behavior is sampled directly there is no need to demonstrate validity (Ebel, 1960) since supposedly no inference from the test data are being made. As will be noted by others during this symposium, this is an inaccurate analysis of the uses of behavioral assessment procedures. Indeed, except in rare circumstances, the behavioral assessor is usually interested in making a number of inferences based on the responses observed during the testing situation.

Without denying the need for various types of validities, I would like to discuss one type of validity, content validity, show its relevance for behavioral assessment and look at methods of establishing

and evaluating the content validity of behavioral assessment procedures. My discussion will focus on the relationship of behavioral assessment to criterion-referenced testing and the application of generalizability theory for both establishing and evaluating the content validity of behavioral tests. I should make it clear, that I am using the word "test" in its broadest sense and include the full range of assessment procedures, from paper and pencil methods to direct observations of behavior in the natural environment.

As pointed out by Goldfried and Kent (1972) and expanded upon by many since, the fundamental assumption underlying behavioral assessment is that the observed behavior is a sample of the behavioral domain of interest rather than a sign of internal attributes. This assumption and the associated interpretations of behavioral assessment data suggest a close parallel between behavioral assessment and criterion-referenced, sometimes called content or domain referenced assessment which shares the same assumption. Criterion-referenced testing, as, distinct from norm-referenced testing, has been one of the most popular and most controversial topics in the field of educational measurement in the last few years (Livingston, in press). It is also becoming increasingly popular in employment settings both as a method of selection and as a method of evaluating employee performance. Although an agreed upon definition is hard to find, criterion-referenced testing is a method of testing designed to estimate a person's level of skill, ability, or achievement in absolute terms. The focus is on what a person can do and on his or her performance in certain specified content domains. This emphasis on skill or ability measurement is

similar to the emphasis many behavioral assessors put on measuring a person's response capabilities where capability is defined as whether or not a person has a particular behavior in his or her behavioral repertoire. The distinction between "response capability" and "response performance," proposed by Wallace (1966) within the context of personality assessment, is similar to the distinction between "response acquisition" and "response performance" proposed by Bandura and Walters (1963). Thus behavioral assessment can be viewed as an application of criterion-referenced testing in at least two ways: 1) both assume that test behavior is a sample; 2) both focus on measuring capabilities and performance in specified content domains. This similarity is important because although the need for content validity is generally recognized and accepted for criterion-referenced testing, the relevance of content validity to behavioral assessment has not been as readily recognized.

Behavioral assessors are often interested in a variety of different domains. A useful way to conceptualize these domains is in terms of universes of generalization. Specifically, as this term is used by Cronbach and his colleagues (Cronbach, Gleser, Nanda and Rajaratnam, 1972) in the context of generalizability theory, the term universe refers to the whole collection of measurements that might have been made. The observed score, be it a frequency count, a sum, or an average, is an estimate of the score that would have been obtained if one had (hypothetically) observed exhaustively from all conditions within the universe. And the question one must ask is to what extent can one validly generalize from the sampled or test

responses across the entire universe of interest. The answer to this question is dependent on the satisfaction of certain requirements of content validity which must be considered in the construction and evaluation of behavioral assessment procedures.

Briefly, the requirements for content validity include first, an adequate specification or definition of the universe of interest, second, representative sampling from the universe and third, specification of the method of evaluating responses and combining them to form a score.

A universe can be defined in terms of facets which alone or in combination refer to the admissable observational conditions. A facet which has received considerable attention in behavioral assessment is that of situations. Although on an ideal level the specification of the universe should include an enumeration of all the situations falling within the domain so that random sampling can be done, in practice this is an impossible task. Goldfried and D'Zurilla (1969) in a seminal article on behavioral assessment and content validity have suggested a two step procedure which closely approximates the ideal of random selection. Briefly, their method involves first constructing a verbal description of the situational universe with a clear specification of the rules for inclusion and exclusion of situations. For instance, rules for inclusion might involve all situations requiring a specified behavioral task such as refusal of unreasonable requests. Rules for exclusion might involve specified settings such as schools if the test is designed for a non-college population. This description is then given to a

random sample of the population of interest who are asked to either generate or record situations fitting the description. A large pool of situations can thus be generated and random sampling can be used to select situations for inclusion in the behavioral test. A more sophisticated procedure would be to group the situations in categories along relevant dimensions, to develop weights for each category, perhaps based on the frequency of occurrence of situations in the category, and then to randomly sample from each category. In essence, this would involve a stratified random sampling procedure in test construction.

If stratified random sampling procedure is used, one runs into other problems in developing a content valid test. Specifically the problem is specifying the basis for stratification. The questions of what are the relevant dimensions for categorizing situations is a complex one and is intimately related to the objectives of the tesing procedure. From a practical point of view, the problem can be thought of as determining on what criteria one should define situations as similar or dissimilar. Depending on the objectives of the assessment, situations might be categorized in terms of settings (e.g. school vs. home), attributes of other persons in the situation (e.g. friends vs. strangers, males vs. females), or any number of other dimensions. At the present time, the specification of relevant situational dimensions for most response classes of interest cannot be done with any degree of precision. What is needed is a body of research specifying the effects of particular situational attributes on specified response classes. Some promising

starts have been made in this direction. For instance, Eisler, Heiser and Miller (1975) have recently obtained evidence that both sex and degree of intimacy of the other person in the situation are relevant attributes in assessing assertive behavior. It is questionable, however, whether it is even possible to arrive at any description of relevant situational attributes applicable to all persons. Bem and Allen (1974), for instance, have provided evidence that situational similarity and consequent behavior are as much a function of the person observing the situation as they are objective characteristics of a situation.

A second facet of interest to behavioral assessors but commonly overlocked in specification of the universe of generalization is that of task requirements or instructions to the subject. Since task instructions can vary from telling the person to just continue with everyday activities in the case of naturalist observations to instructing him or her to try to approach and pick up a snake in a behavioral avoidance test, content validity may apply to one universe of injunctions but not to another. A confusion between the testing universe and the generalization universe is, I think, responsible for some of the problems in establishing criterion validity for some behavioral tests. For instance, often the generalization universe of interest is defined as including typical situations in a specified content domain that a person may encounter in everyday life. If the testing situations are drawn from a universe which is implicitly if not explicitly defined by an injunction to "do your best" or if inhibiting stimuli present in the natural

environment are not represented in the testing situations then it is no wonder that generalization is not found. The test stimuli are not representative of the stimuli in the natural environment. This concern is especially relevant when the objectives during test construction are to estimate response capability but the test is used to estimate response performance. For instance, if one wants to measure whether a person has a response in his or her behavioral repertoire, the universe might be defined in terms of situations where certain demands are made, specified consequences or incentives are included and specified inhibiting stimuli are absent. If there is adequate representation of incentive conditions or directive stimuli there will be content validity for that universe. It would be a mistake, however, to then conclude that the test necessarily has content validity for generalization to typical performance domains.

A similar analysis of the requirements of content validity over other universes could be made. In general, evaluation of content validity of a behavioral test is dependent on the specification of the universe of generalization. Representative sampling may be required for facets of time, behavioral dimensions or tasks, observers, items or any number of other universe definitions. Hopefully, the application of generalizability theory to behavioral assessment and the associated requirements for content validity will alert the behavioral test developer to the need for both adequate universe specification and representative sampling of conditions within the universe.

An additional issue involves how one might empirically demonstrate the content validity of a particular behavioral test. Cronbach

(1971) suggests the construction of parallel tests using the same universe definitions, sampling rules etc. The correlation between responses from the two tests would then serve as an index of content validity. Another approach, to assess the adequacy of the classification of conditions, would be to conduct generalization studies. Theoretically, there should be generalization across each specified category or strata within the universe of interest. If this turns out to be the case, then one can be reasonably sure that random sampling from within each category will result in a representative sampling of the content domain of interest.

In conclusion, I would like to briefly discuss the issue of idographic vs. nomothetic test construction. This is an important issue for at least two facets of generalizations: situations and behavioral or task dimensions. With respect to situations, although I have emphasized a nomothetic approach to defining the universe it is equally possible to develop, categorize, and sample situations on an idiographic basis. In fact, the common clinical practice of having a client keep a record of all encountered situations of a certain type is an example of this. To the extent that the client is observed in a representative sample of these situations the test is content valid. The same points can be made for the facet of behavioral dimensions. The requirements of task analysis to generate a list of behaviors to be observed related to a specified task domain can be done on either a nomothetic or idiographic level. Thus the specific response classes making up a domain as well as the discreet behaviors within each class may be defined by the test developer from either an empirical analysis of the task requirements in a particular

domain or on the basis of theoretical considerations. On the other hand, the universe may be defined by an analysis of which discreet behaviors covary for a particular individual and thus form a category or class within which one might randomly select behaviors for observation. It is not, however, necessary nor even desirable that responses covary across the entire universe of interest.

Whether it is possible to develop useful behavioral assessment procedures on a nomothetic level is primarily an issue of both the objectives of the testing and the degree to which generalization is sought. At present, there is a great need to do more research to define those universe specifications which will allow the development of content valid behavioral tests.

In summary, both criterion-referenced testing and behavioral assessment share the basic assumption that test behavior is a sample rather than a sign. In addition, the focus of both types of assessment are on response capabilities and performance in specified content domains. Although content validity has been traditionally recognized as essential to criterion-referenced testing, it has not been as generally recognized as relevant to behavioral assessment. The relationship between the two types of assessment clarifies the importance of content validity for behavioral assessment. Questions of content validity can be usefully conceptualized from the point of view of generalizability theory. The relevant question then becomes: Are the test items a representative sample of the universe to which results are going to be generalized? Finally, content validity is as important in assessment procedures developed on an idiographic level as it is in those developed nomothetically.

References

Bandura, A., & Walters, R. H. Social learning and personality development. New York: Holt, Rinehart & Winston, 1963.

Bem, D. J., & Allen, A. On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. Psychological Review, 1974, 81, 506-520.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational Measurement. Washington: American Council on Education, 1971.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Ebel, R. Must all tests be valid? American Psychologist, 1960, 15, 546-553.

Eisler, R. M., Hersen, M., & Miller, P. M. Situational determinants of assertive behaviors. Journal of Consulting and Clinical Psychology, 1975, 43, 330-40.

Goldfried, M. R., & D'Zurilla, T. J. A behavioral-analytic model for assessing competence. In C. D. Spielberger (Ed.), Current Topics in Clinical and Community Psychology, Vol. I. New York: Academic, 1969.

Goldfried, M. R., & Kent, R. Traditional vs. behavioral assessment: A comparison of methodological and theoretical assumptions. Psychological Bulletin, 1972, 77, 409-420.

Livingston, S. A. Psychometric techniques for criterion-referenced testing and behavioral assessment. In J. D. Cone & R. P. Hawkins (Eds.),

Behavioral assessment: New directions in clinical psychology. New York: Brunner-Mazel, in press.

Mischel, W. Personality and assessment. New York: Wiley, 1968.